Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation

**Mehmet Kayaalp, MD, PhD, Allen C. Browne, MS,
Pamela Sagan, RN, Tyne McGee, BS, Clement J. McDonald, MD
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD**

Abstract

The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. We have been manually annotating clinical text since 2008 in order to test and evaluate an algorithmic clinical text de-identification tool, NLM Scrubber, which we have been developing in parallel. Although HIPAA provides some guidance about what must be de-identified, translating those guidelines into practice is not as straightforward, especially when one deals with free text. As a result we have changed our manual annotation labels and methods six times. This paper explains why we have made those annotation choices, which have been evolved throughout seven years of practice on this field. The aim of this paper is to start a community discussion towards developing standards for clinical text annotation with the end goal of studying and comparing clinical text de-identification systems more accurately.

1. Introduction

The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. The Rule indicates 18 pieces of personally identifiable information (PII) that need to be de-identified in order to protect patient privacy. Our particular interest in the Privacy Rule is to use it as our guideline for preventing unintended privacy breaches during the secondary use of patient health information for clinical research. Although it is very clear what each piece of PII is, conceptually, it may not be so straightforward when we try to do so manually in a clinical narrative report. The Privacy Rule has been designed mainly with the structured tabular data in mind. When we attempt doing the same with text, we are faced with a number of difficulties that arise due to the nature of English, or any other natural language.

For example, the Privacy Rule requires de-identification of personal names but does not say anything about personal name initials (e.g., JFK). While U.S. District Courts impose restrictions on the use of personal names of minors and require their names in all hearing transcripts to be de-identified by converting them into personal name initials,[1-3] the Office of Civil Rights of the Health of Human Services interprets the Privacy Rule in such a way that equates personal name initials with personal names.[4]

It is unreasonable to expect from any piece of legislation or set of rules to spell out every imaginable version or combination of entities that could occur in a clinical narrative report or patient record. Our approach is to take the Privacy Rule as a model, interpret its language, understand its aim, and in ambiguous cases, make decisions whether we ought to de-identify those particular pieces of information in order to fully comply with the Privacy Rule to the best of our abilities.

To this end, we have been developing annotation guidelines, which basically are a compendium of examples, extracted from clinical reports, to show what types of text elements and personal identifiers need to be annotated using an evolving set of labels. We started annotating clinical text for de-identification research in 2008, and since then we have revised our set of annotation labels (a.k.a. tag set) six times. As we are preparing this manuscript, we are working on the seventh iteration of our annotation schema and the label set, and will be making it available at the time of this publication.

Although the Privacy Rule seems pretty straightforward at first glance, revising our annotation approaches so many times in the last seven years is indicative of how involved and complex the task is. We don't believe that publishing the guidelines would suffice by themselves, since the guidelines only tell what needs to be done. In this paper, we try to address not only what we annotate but also why we annotate the way we do. We hope that the rationale behind our guidelines would start a discussion towards standardizing annotation guidelines for clinical text de-identification. Such

standardization would facilitate research and enable us to compare de-identification system performances on an equal footing.

Before describing our annotation methods, we provide a brief background on the process and rationale of manual annotations, discuss personally identifiable information (PII) as sanctioned by the HIPAA Privacy Rule, and provide a short overview of approaches of how various research groups have adopted PII elements into their de-identification systems. We conclude with Results and Discussion sections.

2. Background

Manual annotation of documents is a necessary step in developing automatic de-identification systems. While de-identification systems using a supervised learning approach necessitate a manually annotated training sets, all systems require manually annotated documents for evaluation. We use manually annotated documents both for the development and evaluation of NLM-Scrubber. [5-7]

Even when semi-automated with software-tools,[8] manual annotation is a labor intensive activity. In the course of the development of NLM-Scrubber we annotated a large sample of clinical reports from the NIH Clinical Center by collecting the reports of 7,571 patients. We eliminated duplicate records by keeping only one record of each type, admission, discharge summary etc. The primary annotators were a nurse and linguist assisted by two student summer interns. We plan to have two summer interns each summer going forward.

These annotators used NLM's Visual Text Tagging tool, VTT. VTT allows annotators to select and annotate strings of text by swiping the cursor over them and choosing a tag from a pull-down list of annotation labels. The application displays the annotation with a distinctive combination of font type, font color and background color. Tags in VTT can have sub-tags which allow the two dimensional annotation scheme described below. VTT saves the annotations in a stand-off manner leaving the text undisturbed and produces records in a machine readable pure-ASCII format. A screen shot of the VTT interface is shown in Figure 1. VTT has proven helpful both for manual annotation of documents and for displaying machine output. As an end product the system redacts PII elements by substituting the PII type name (e.g., [DATE]) for the text (e.g., 9/11/2001), but for evaluation purpose tagged text is displayed in VTT.

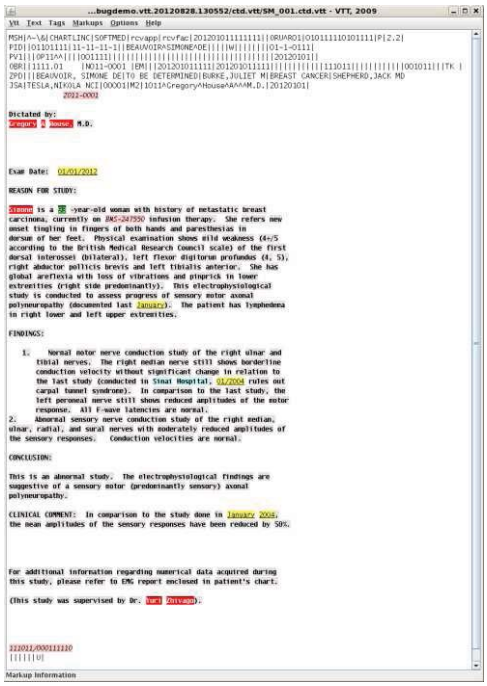

Figure 1. VTT Window Showing a Hypothetical Annotated Report

The Privacy Rule guidelines published by the Office of Civil Rights of the Health and Human Services (HHS) say that "The importance of documentation for which values in health data correspond to PHI, as well as the systems that manage PHI, for the de-identification process cannot be overstated." [4]

Most studies in the area of automatic de-identification only indicate the set of PII items they redact in the course of de-identification. In their review, Meystre et al.[9] mapped the personal identifiers that were de-identified by the 18 different systems into seven categories: patient names (or both patient and provider names), ages greater than 89, geographical locations, hospitals and healthcare organizations, dates, contact information (phone numbers, pager numbers, fax numbers, and e-mail addresses), and IDs (Social security number, medical record number, driver's license number, and other identifiers).

The i2b2 challenge uses a list of eight identifier types: patient, doctor, location, hospital, date, ID, phone, and age.[10] Studies based at the Veteran Administration Health System treat a different list of items more in line with their particular needs adding four non-PII categories of clinical eponyms to the i2b2 list.[8] They annotate medical procedure names, medical device names, disease names, and anatomic structures. By supplying the annotations of significant clinical information, they could train their supervised learning system so that it could recognize these entities in text and then evaluate how well the system performs in preserving such clinical information at the end of the de-identification process.

Separating doctor names from patient names is another move in this direction. Although the names of doctors and other medical personnel are not PII, they could not be distinguished with a high level of confidence by automatic de-identification systems and might be redacted inevitably.

In most cases, the details of the annotation scheme are not published. The i2b2 efforts publicly provide their corpus with a data use agreement.[10] This is only possible through an automatic de-identification process followed by an extensive multi-round process of manual validation by human experts. In the resulting corpus the PII elements are substituted with pseudonyms, a surrogate text that looks like the original.

This emphasis on annotation of PII alone overlooks a need for more elaborate annotation, including non-PII items and specific sub-parts of PII items that contribute to evaluation and error analysis and the need to explicitly publish the guidelines used to annotate documents. This paper discusses the annotation scheme we use in the NLM Scrubber project detailing the annotation guidelines and the reasoning behind the decisions that led to this annotation schema.

3. Methods

As mentioned in the previous sections, we have been annotating clinical text since 2008. Our main goal in this effort has always been to develop a set of standards[7] so that we can evaluate the performance of our clinical text de-identification system, NLM Scrubber.[5, 6] Both our annotation and de-identification studies have been influencing and informing each other. While the availability of a standard text has been helping us to test new ideas and monitor the de-identification performance of our system as we modify existing modules, evaluation of the updated de-identification system has also made the shortcomings of our annotation methods explicit. In the following subsections, we describe our annotation methods and explain why we annotate the way we do.

3.1. Annotations on two dimensions

We perceive the annotation space in two orthogonal dimensions. The first dimension denotes personal identifiers. We established a total of 12 personal identifier categories: Address, Personal Name, Personal Name Initials, Organization, Occupation, Telecommunication, Date, Age, Time, Numeric and Alphanumeric Identifiers, Personally Identifying Context, and Role.

The second dimension is personhood, which associates the identifier with an identity. We define 5 personhood categories: Patient, Relative, Employer, Provider, and Other. For example, we would annotate the word "John" in the following two dimensions: It is a personal name and may denote (say) the patient. If the latter is true, we would use the following label PersonalName::Patient. If "John" is the name of the health care provider, we would label it PersonalName::Provider.

We use the personhood category Relative broadly, which includes family members as well as the members of the household of the patient—the Privacy Rule mentions them separately. Given that a family member mentioned in a clinical report is frequently a household member as well, categorizing them separately would be problematic, since we would have to annotate the same word with two distinct personhood labels. Although technical challenges are not insurmountable, it would be conceptually too complex for the annotators to distinguish whether the family member mentioned in the clinical text was also living with the patient in the same house.

Although the Privacy Rule dictates that personal identifiers of the patient's employer must be de-identified, it does not clarify what constitutes an employer. It could be the owner, president, or the CEO of the company. Could it be the supervisor of the patient? How about their supervisors? In many workplace accident cases, the patient is accompanied to the health care facility by a co-worker. In a re-identification attempt, the co-worker's identity could be linked to the company and through which, indirectly, to the patient; thus, we use the personhood category Employer to annotate all types of co-workers and supervisors of the patient.

The Provider category denotes every type of healthcare professional who takes part in the health care of the patient. Note that information about the provider was not defined by the Privacy Rule as PII. We use the category Other to denote other personhood identities that are not patients, relatives or providers and there is no apparent method to link that particular person or personal identifier to the patient. For example, we annotate the word Obama in "the patient cited Obama as our president" with the label PersonalName::Other. Disclosures of identifiers associated with Provider or Other usually do not pose any significant privacy risk to the patient, since they are not directly linkable to the patient.

How should we annotate girlfriend, partner, and neighbor? We annotate partner as Relative, since it may indicate some kind of formal union and/or household membership, and can be linked to the patient. We use the label Other for friends and other informal relations who may not be linked to the patient directly and as easily as a household member—in the age of social networks, we are not sure how long this assumption would be holding! Although neighbor seems fitting to the label Other at the very first glance, the neighbor information is actually akin to that of the household member, since their residence information could be identifying the address of the patient; thus, we annotate it as Relative.

By reserving the label Other for information that cannot be linked to the patient directly (or indirectly) and by not using it for sensitive information such as information about neighbors, we may prevent significant complications with respect to the evaluation of the de-identification system in case of any unintended disclosure.

In the following subsections, we discuss 12 personal identifier categories, what subcategories, if any, they consist of and how they are related to identifiers mentioned in the HIPAA Privacy Rule. Some entities in these categories may not be personal identifiers. In those cases, we discuss why we chose to introduce and annotate them.

3.2. Address

The Address category comprises a number of entities such as street name, number and types. Table 1 shows which labels we use to annotate such entities. A mention of address may contain a subset of these entities.

Table 1. Address Labels

Label	Entity	Example
Street	Street name	Pennsylvania Ave
Location	Street number, apartment, suite or office number; floor or room number inside an office building, hospital or clinic including a bed number, P.O. Box	Station 10-Room 33-A
Building	Building name	Woodward Building
City	Village, town or city	Bethesda
County	County	Montgomery County
State	State, US district, territory, province or region	D.C. Metro Area, Guam, East Coast, Alberta, Western Pennsylvania
Country	Country	United States Mexican-American
Zip	Five or nine digit US ZIP code or foreign postal equivalent	20894-3828, SW1A 2AA

Why do we use eight different address labels, instead of using a single label, to annotate all address tokens? Using a single, common address label sounds quite practical at the first glance, esp. during the annotation process. However, if one needs to assess the performance of a de-identification system that may inadvertently reveal some address information, uniform address labels would be very inadequate for estimating the level of risk to the potential breach

of patient privacy. Note that revealing certain address elements, e.g. a rare street name and number, could pose significantly more risk than revealing more common or widely shared address elements such as an apartment number or name of the city where the patient resides.

HIPAA Privacy Rule makes a distinction between different types of address information. The Privacy Rule states that information about all geographic subdivisions smaller than state, except the first two digits of the zip code, must be de-identified. The third digit of the zip code can be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data. In other words, the Privacy Rule indicates certain address tokens are more informative than others in identifying an individual. If we visualize the address elements on a line ordered from the most granular or specific elements (such as street name and number) towards the most widely shared element (i.e., country), the Privacy Rule puts the threshold between County and State.

If the user intends to fully de-identify patient data, then s/he needs to use the above threshold. However, the Privacy Rule also offers a lower threshold in its Limited Data Set provision, which allows the user to preserve city and town information as long as such information is necessary for the study and the user signs a data use agreement with the provider of the data.

These two thresholds divide the address elements into three parts: If using the Privacy Rule, (A) information more specific than town or city needs to be eliminated under any circumstances; (B) state and country information can be preserved even in a fully de-identified set of data; and (C) information whose specificity lies between these two thresholds that can be preserved only within the boundaries of the Limited Data Set provision. In other words, one needs to use at least three distinct labels to differentiate these three parts in an address. Furthermore, a separate label for ZIP codes is also necessary since ZIP code information crosses these two boundaries.

We could merge State and Country labels into one State/Country label and merge City and County labels into City/County label but we chose not to do so for various reasons. The first reason is practicality—annotating these four types of address elements separately does not impose undue burden onto our annotators. Distinguishing these labels can also be useful under certain situations. For example, in an epidemiologic study in which preserving county information may be necessary and sufficient, de-identifying city information could better protect patient privacy. Unless the user requires state information, NLM Scrubber de-identifies it by default. Although HIPAA does not sanction state information to be de-identified, we choose to do so, since many re-identification algorithms rely on address information and the more unnecessary address information we could de-identify, the more difficult the re-identification would be. Since it is usually difficult to distinguish country of residence from the country of origin, we annotate the residence of a foreign national and the country of the origin of an individual (e.g., “40yo Ethiopian man”) with label Country. We believe a de-identification system ought to preserve ethnicity and country of origin information, since some diseases are more prevalent in certain groups and geographical locations globally.

We distinguish three distinct address elements below the town or city level: Street name, location information such as house or apartment numbers, which further qualify the local address, and building name. Note that inadvertently revealing a street (i.e. house) number without disclosing the street name would not jeopardize privacy of the patient—it would be truly nonspecific. Revealing a street name without the street number however poses a more serious risk, especially if the street name is not a very common one. In that scenario, the privacy risk would be inversely proportional to the size of the household population on all streets with that street name in the country. If on the other hand, the de-identification system inadvertently discloses both street name and number, re-identifying the individual along with age and gender information may not be too difficult. We separated building names from street and location categories, because a building name alone can be more informative than either of them. Note that a building name is at least as informative as the combination of both the street name and the street number. Since it is not customary to name residential units in the US, a building name, as rare as it is, could be quite identifying.

3.3. Personal Names and Personal Name Initials

The Privacy Rule states that names (of the individual or of relatives, employers, or household members of the individual) should be removed. If one has a tabular data where the columns are well defined, it would be easy to distinguish personal names from other identifiers. But when the data in question is text, the seemingly obvious de-identification task can be quite complicated. For example, questions like “would it be okay to leave single letter middle initials intact?” or “would personal name initials constitute a name?” can be difficult to answer.

Clearly, personal name initials like JFK are not as revealing as corresponding full names. In fact, converting names into initials is a widely used practice to protect identities of the minors in reports of the court hearings.[1-3] So, we

categorize personal name initials separately from personal names. According to the Office of the Civil Rights, however, personal name initials are considered as personal names and ought to be de-identified.[4] We reserve personal name initials only for the full set of name initials (i.e., when first, middle, and last names are initialized altogether as in JFK) but annotate middle and/or first name initials, as in "James T. Kirk" or "J.K. Rowling," as parts of the personal names.

Although we annotate suffixes such as Jr. and Sr. as parts of personal names, we do not extend it to professional and academic titles, for some of which we use the label Occupation.

3.4. Occupation and Organization

Occupation information is not one of the 18 pieces of PII, sanctioned by HIPAA, to be de-identified. However, especially if it is a rare occupation (e.g., clinical computational linguist, Supreme Court justice), the information may be used to re-identify the patient. Up to date, we have not come up with an easily implementable annotation method to differentiate rare occupation information from the common ones. We have to separate the wheat from the chaff for each piece of occupation information at the evaluation phase of our de-identification studies. Note, however, the personhood dimension that we introduced in this paper for the first time (see Section 3.1) can be helpful when occupation information is associated with Provider or Other, which usually would not pose any privacy risk to the patient.

Most professional titles indicate the occupation of the person. Although we annotate provider occupations (e.g., dermatologist) whenever it is explicitly stated in the text, we have not been annotating their titles (e.g., Dr., M.D., etc.) due to their sheer number of occurrences and the difficulty that it would impose on our annotation team. We are currently studying the feasibility of the issue in a pilot.

We also annotate past occupation information but not the future ones. The former can be linked to the patient but the latter (e.g., "the patient plans to open a car dealership") is mostly hypothetical. Similarly, we do not annotate hobbies as occupations since they would rarely be unique and linkable to the patient. In such rare scenarios, however, we have other methods to employ (see Section 3.7).

Occupation (e.g. a cook) does not specify the employer like where the person works (e.g., "… at Acme Restaurant"), but sometimes, they are very closely linked together. For example, "he is an Army Master Sergeant," where we annotate Army with label Organization::Employer and Master Sergeant with Occupation::Patient or Occupation::Relative, depending on whom "he" denotes. If the title were Admiral, for which we would use label Occupation::Patient, it would also implicitly reveal the employer's organization, Navy.

We reserve the personhood label Employer only for the patient's employer and do not extend it to the employer of the relative, since there is no apparent direct link from the employer to the patient. In the example, "The patient's mother is a math teacher at Takoma Park Middle School," math teacher is Occupation::Relative and Takoma Park Middle School is Organization::Relative. Between the school and the patient, there is two degrees of separation, which is implied by the label Organization::Relative—the linkage for re-identification is possible but the link is weaker than the link between the patient and their employer.

Although we do not annotate hobbies, we do annotate organizations that individuals can be associated with (e.g., "the patient is a member of the Rotary Club" or "…presented his findings during the AMIA Symposium last year").

3.5. Age, Date and Time

Similar to category Address, Age and Date are categories, each of which comprises multiple labels. By mandating that ages over 89 be de-identified, HIPAA separates age into two categories: (1) ages 90 and above are considered PII, which we annotate with label AgePII, and (2) ages that are below 90, which HIPAA considers as non-PII. We split the second group into two additional separate groups: (2A) ages that are mentioned as whole numbers, which we annotate with label AgeNPII, and (2B) ages that are mentioned as fractions of whole years (e.g., "Patient is a 4 and 11/12 month old boy"), which we annotate with label AgeFraction.

Without an anchor to a fixed date AgeFraction is not very useful to re-identify the patient; thus, it should be considered as non-PII. However, it is possible that a de-identification system might miss a mention of the report date, which, along with the age information in fractions (e.g., "he will be 11 months old in two days"), one may be able to identify

the birth date of the patient. In other words, label AgeFraction could pose privacy risk only in conjunction with an inadvertently revealed full-date within the text.

If the patient's current age is 90 or older and the narrative report provides indirect reference to the patient's age such that re-identification can be done through a simple arithmetic (e.g., "Twenty years ago, at the age of 75, he had an ischemic attack"), we would annotate the earlier age references (i.e., 75 in the example above) as AgePII as well.

We do not annotate other "age" types such as gestational age, bone age (unless identical to the chronological age), school grade level (10th grade) or age periods such as teenage, middle-aged, etc., since they are not as identifying as chronological age found in formal records.

The category Date comprises six labels: Year (e.g., 2001), Month (e.g., September), Day (e.g., 11th), DayOfWeek (e.g., Tuesday but not Tuesdays), SpecialDay (e.g., 9/11, Hurricane Sandy, Katrina, Cinco de Mayo, New Year), and Period (e.g., flu season, Monsoon, Ramadan, winter, second trimester).

We annotate not only those special days that are fixed in history such as Pearl Harbor, 2008 Market Crash but also those special days that occur every year such as New Year, whose exact dates can be construed when combined with year information, which taken alone is not PII under HIPAA. We also label personal special days such as birthday or Bar Mitzvah, not only due to potential privacy concerns as they may be available from external sources, but also due to their potential importance in reference to other events in the narrative text.

We use the label Period to annotate any time period longer than a day of which begin and end dates are not explicitly stated. We use this label to annotate periods in the patient's medical history such as pregnancy, puberty, hospitalization period, and menstruation as well as calendar periods such as early 2001 or in the 90s. Most age references in the medical history are periods. For example, "when the patient was 5 years old …" or "spoke at 5–5 ½ years old". Note that age references in these examples do not denote the patient's current age but if such age references in the past reveal that the patient's current age is 90 or older, we would have to use label AgePII instead.

If a period of two days or longer is described in terms of an interval or a range with explicit begin and end date identifiers (e.g., 1995–97, between next Tuesday and Friday), we separately annotate begin and end points with the appropriate date label. If it is an age range, we label each age separately. In the example, "had hearing loss from 85-97 years old", we annotate 85 with AgeNPII and 97 with AgePII.

Recall that we define the Period as a subcategory of date; therefore, we use it only if the period can be stated relative to a date. In example, "when the patient was 5 years old", we perceive a period of one year, starting 5 years after the birth date. If the period is stated using terms like last year, last month, last week, and last weekend, the period is defined relative to the date of the report. We do not annotate (hence do not use the label Period) cyclical temporal references such as daily, Tuesdays or every Tuesday or other temporal references described in sequence of events without any apparent date to anchor (e.g., "completed 2 weeks of antibiotics").

We annotate last Christmas or Christmas last year as SpecialDay since the terms last and last year further qualify the special day, but when the year is explicitly stated as in Cinco de Mayo 2000, we annotate Cinco de Mayo as SpecialDay and annotate 2000 as Year, because in this example, the date term refers to a full date May 5, 2000.

We do annotate time of the day using the label Time, but we also believe that it is too general to link to the patient for re-identification. Since we do not classify it under the Date category, we do not annotate time periods within a day as Period (e.g., noon-4:30pm); instead, we use label Time to annotate noon and 4:30pm, separately.

3.6. Telecommunication and Alphanumeric Identifiers

Telecommunication identifiers are the most straightforward and the least ambiguous identifiers since they are well defined engineering objects. Of the 18 personal identifiers defined by the HIPAA Privacy Rule, five of them are telecommunication identifiers to be de-identified: telephone numbers, fax numbers, electronic email addresses, web universal resource locators (URLs), and Internet protocol (IP) address numbers. As new telecommunication modes and media emerge, new telecommunication identifiers (e.g., Twitter usernames such as @BarackObama) appear, but they too are covered by the last (18th) catch-all identifier of the Privacy Rule: "any other unique identifying number, characteristics, or code" must be de-identified.

Numeric and Alphanumeric Identifiers consist of four labels: MedicalRecordNo, ProtocolNo, HealthRecordID, and AlphanumericID. The first two are very specific identifiers denoting medical record and protocol numbers,

respectively. Medical record number is one of the 18 HIPAA identifiers. Since protocol numbers are very important entities for clinical researchers, who are the intended users of NLM Scrubber, we annotated them separately. We use the label HealthRecordID for all other alphanumeric identifiers issued by health care and insurance providers uniquely to the patient; e.g., hospital account number, health plan beneficiary number and lab specimen number. MedicalRecordNo, ProtocolNo and HealthRecordID are almost always associated with the patient— only in a handful of cases we did observe mentions of such identifiers for the relatives.

We use the label AlphanumericID for all other numeric and alphanumeric identifiers that are not issued by the provider, including those five identifiers defined by the Privacy Rule: social security number, account numbers, certificate / license numbers, vehicle identifiers, and device identifiers. Note for hospital account numbers we use the label HealthRecordID.

Sometimes, names of some lab materials and experimental drugs may contain some numbers (e.g., drug 123-ABC or instrument QRS-40). We do not annotate such health information, as they are neither unique to the patient nor personal identifiers.

3.7. Personally Identifying Context

So far, we discussed how we annotate entities that were mentioned in the HIPAA Privacy Rule along with a few other closely related entities, some of which can be PII in certain contexts. We are aware of the fact that due to the intricacies of natural languages, it is possible to specify a context in which the person could be identified indirectly such that no labels we discussed so far would be appropriate to use. In those cases, we label the tokens with PIC, denoting Personally Identifying Context.

In the hypothetical example, "received his injuries while he was reporting from Tahrir Square", we would annotate reporting with label Occupation::Patient and Tahrir Square with PIC::Patient, since the latter would provide context so specific that along with the occupation information would probably identify the person directly.

In the example "the patient was deployed to Iraq", the reader may presume that the patient is in the military, but the deployment to Iraq is not an occupation—equipment could be deployed to Iraq as well as other types of personnel such as reporters could be deployed to a war zone. However if the example provides an occupational context that is so specific that it might tighten the circle of potential candidates, we would label those tokens as PIC. But in this example, even if we presume that the context alludes that the subject is a military person, the circle of military personnel remains too broad to label the phrase as PIC.

3.8. Role

In order to associate a personal identifier with a person, automatic de-identification system needs to recognize a reference to that person. We define such a reference as Role, which can denote the patient, mother, father, daughter, supervisor, physician, boyfriend, and others. We annotate those roles in order to evaluate and monitor our system's performance. Although they too are roles, we do not annotate pronouns such as he, she, him, hers, their, themselves etc. We use the label Role only if no other label is suitable for that annotation. For example, if the provider's occupation is more specific than the role of physician or nurse, such as cardiologist or physical therapist, then we annotate it as Occupation. If the reference specifies a personally identifying context, instead of using the label Role, we would annotate it as PIC.

The role information is quite important in the context of the deceased patient records as well,[11] because even though health records of the deceased patient may not constitute protected health information, health information of their living relatives does. Fortunately, such information is quite rare. Recognizing such roles in the narrative reports of the deceased helps prevent such privacy breaches.

4. Results

Our annotation label set and methods of annotating text elements that we described in this paper are the results of the seven years long evolution of annotation, de-identification, and evaluation. By defining the annotation labels on two dimensions and associating identifiers with personhood, Patient, Relative, Employer, Provider, and Other, we can easily stratify the importance of text elements in terms of high, medium, low, and no privacy risks.

We divided some identifier categories such as Address into subcategories, each with a distinct label. Even though some information (e.g., house or street numbers labeled with Location) seem more granular or specific than others (e.g., town labeled with City), inadvertently revealing them would pose little or no privacy risk; however such identifiers (e.g., house number and street name) become very significant only if they are revealed in combination with certain other elements of the same category (e.g., house number and street name together). The same is true for the subcategories of Date; i.e., day, month, or year information alone has no significance until they are revealed together.

The newly introduced special subcategories and associated labels such as Period, SpecialDay, and AgeFraction enrich our label set and provide clarity and direction to our annotators when faced with non-standard and borderline cases. For example, "at age 3, the patient started…" may seem to contain a piece of information about the patient's age at the very first glance, but "age 3" actually outlines a period in the medical history of the patient and does not identify how old the patient currently is. In short, these new labels yield a corpus with more accurate annotations.

Personally Identifying Context labeled with PIC is a very important new category since we no longer need to say "There could be some information identifying the patient indirectly without using any explicit PII elements in this report." Now, if we encounter such information, we have the tool to annotate it.

5. Discussion

In this paper, we introduced a new annotation schema that extends the identifier elements of the HIPAA Privacy Rule. In this schema, we annotate text elements on two dimensions: identifier type and personhood denoted by the identifier. The personhood can take one of the following type values: Patient, Relative, Employer, Provider and Other. We extended identifier types both in terms of scope and granularity.

Our annotation label set is based first and foremost on the PII elements defined by the HIPAA Privacy Rule. However, being aware of other annotation efforts, we tried to design a broad spectrum of annotation labels so that we can establish a common ground for our community. Standardization of annotation schemas is a very important goal that we all should strive for; otherwise, an effective evaluation and comparison of our study results would be too difficult. We believe this is the first step towards that ambitious goal.

The concepts and annotation methods defined and described in this paper could be best understood if studied along with a number of good examples. We are currently working on finalizing our annotation guidelines containing a rich set of examples most of which are extracted from actual reports. The guidelines will be publicly available by the time of this publication at http://scrubber.nlm.nih.gov.

Acknowledgements

We are grateful to Brett South, Guy Divita and their colleagues for sharing with us the annotation guidelines used in their research at the University of Utah and the VA Salt Lake City Health Care System.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Competing Interests

The first author receives royalties from University of Pittsburgh for his contribution to a de-identification project. NLM's Ethics Office reviewed and approved his appointment.

References

1. Hanna J. Some Supreme Court Rule 138 privacy provisions delayed until 2015. Illinois Bar Journal 2015;102(2):62.
2. U.S. Courts District of Idaho. Transcript Redaction Policy & Procedures, 2014. URL: http://www.id.uscourts.gov/district/attorneys/TranscriptCourt_Reporter.cfm. Accessed on 3/6/2015.
3. U.S. District Court Southern District of California. Electronic Availibility of Transcripts -- Redaction Procedure, 2008. URL: https://www.casd.uscourts.gov/Attorneys/SitePages/Transcripts.aspx. Accessed on 3/6/2015.

4. Office of Civil Rights. Guidance Regarding Methods for De-idnetification of Protected Health Information in Accordance with Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. In: Services USDoHaH, editor, 2012.
5. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The Pattern of Name Tokens in Narrative Clinical Text and a Comparison of Five Systems for Redacting them. J Am Med Inform Assn 2013.
6. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. Proceedings of the Annual American Medical Informatics Association Fall Symposium 2014.
7. Browne AC, Kayaalp M, Dodd ZA, Sagan P, McDonald CJ. The Challenges of Creating a Gold Standard for De-identification Research. Proceedings of the Annual American Medical Informatics Association Fall Symposium 2014.
8. South BR, Mowery D, Suo Y, Leng JW, Ferrandez O, Meystre SM, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform 2014;50:162-72.
9. Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology 2010;10(1):70.
10. Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. J Am Med Inform Assn 2007;14(5):550-63.
11. Huser V, Kayaalp M, Dodd ZA, Cimino JJ. Piloting a Deceased Subject Integrated Data Repository and Protecting Privacy of Relatives. Proceedings of the Annual American Medical Informatics Association Fall Symposium 2014.